Personalization of AI-based Distance To Empty prediction model

Kihyung Joo
Hyundai Motor Company
Seoul, South Korea
steric5@hyundai.com

Lina Kim
Hyundai Motor Company
Seoul, South Korea
kamooie@hyundai.com

Abstract—It is an important factor in electric vehicles to show customers how much they can drive with the energy of the remaining battery. If the remaining mileage is not accurate, electric vehicle drivers will have no choice but to feel anxious about the mileage. If the remaining mileage to drive is wrong, drivers may not be able to get to the charging station and may not be able to drive because the battery runs out. This study proposes a more advanced model by predicting the remaining mileage based on actual driving data and based on reflecting the pattern of customers who drive regularly. The basic model is a linear regression model, and the advanced model is a Bayesian linear regression model. In order to improve performance, the driver’s regular driving pattern is recognized in advance before driving and it is reflected in the remaining driving mileage model. The actual driving log is used for the dataset. It can be seen that the performance of the model in this study is improved 10% better compared to the existing remaining driving mileage.

Keywords—Distance To Empty(DTE), Linear regression, Bayesian Linear regression, DBSCAN, Electric vehicle

I. INTRODUCTION

The electric vehicle market continues to develop by strengthening regulations on environmental protection in each major developed country. It has long been a paradigm shift from internal combustion engines using fossil fuels to electricity-based power sources. Electric vehicles have better initial torque, less vibration, and very low maintenance cost for fuel use compared to internal combustion engines. In addition, the number of parts is few compared to the internal combustion engine, so there are few breakdowns, and the repair cost is not much. However, when customers purchase electric vehicles, the main considerations are the price of the vehicle, the All Electric Range(AER), and the charging time. Vehicle prices become reasonable through government grant and continuous technology enhancement to reduce costs, and driving mileage is increasing due to improved battery energy efficiency and increased battery capacity [1]. Charging time also become gradually improved thanks to expansion of charging infrastructure and development of charging technology. However, the driving mileage that customers feel does not satisfy them yet. The customers consider the time when they charge vehicles batteries through Distance To Empty(DTE) and State Of Charge(SOC) shown on the cluster like Fig. 1. EV6 is the electric vehicle in KIA. SOC is inconvenient for drivers to recognize remaining distance, DTE shows more realistically how much distance can be driven in near future. The reason why DTE is important in electric vehicles is that unlike internal combustion engine vehicles, electric vehicles should pay more attention to charging time and charging location. In the case of internal combustion engine vehicles, the gas station is very convenient to refuel, and the maximum fueling time is sufficient to be 5 minutes. However, electric vehicles have a very limited charging location, and charging time takes about 30 minutes for fast charging, so drivers are bound to be uncomfortable. Therefore, if the DTE shows accurately the remaining mileage, drivers can feel less frequent charging or range anxiety considering their future driving plan and a charging station [2].

The DTE model algorithm currently calculates the remaining mileage by calculating the available energy of the current battery and the expected fuel economy of the vehicle. In addition, in order to reflect the past driving pattern, it is corrected by considering the past fuel economy(km/kWh) which is from learning parameter, and the current fuel economy which is current driving preference.

The currently implemented DTE does not take into account regenerative braking, road conditions like slope, road type, or traffic jam and etc.. In addition, the model itself does not reflect each individual's driving characteristics or driving conditions. If it is linked to the situation outside the vehicle and the driving conditions of the driver, more accurate driving mileage prediction is possible. This proposed model is devised to predict the possible remaining distance in consideration of SOC charging parts, such as regenerative braking, road conditions, and when to drive. Additionally, the model is finally implemented as the algorithm that can be embedded within vehicle ECU(Electrical Control Unit). Through this, it
aims to develop a personalized model for predicting the remaining mileage in consideration of dataset from the vehicle and driver’s driving conditions.

The main characteristics of the AI-based DTE prediction model through this study are as follows.

1. Select the model that is better than the accuracy predicted by the existing DTE prediction model.
2. Simplified AI model enables in-vehicle learning and implements a model that does not have problems during computing
3. Implementation of the model that enables online learning according to changes in vehicle conditions (battery aging, vehicle friction, etc.)
4. Using driving history, determine what past driving pattern is similar to the current driving pattern and improve the predicted DTE model
5. Implementing algorithms that are good at computational time and resource usage to embed within vehicle ECU
6. verify performance robustness compared to the vehicle’s actual vehicle

Using various machine learning models, a high-performance machine learning model was selected with dataset obtained from real-road driving and driver’s driving data as training data and testing data.

II. RELATED WORK

In order to predict the remaining driving distance, there are two solutions. One is an energy-based DTE model [3], and the other is a mileage-based DTE model [4]. The energy-based DTE model uses a physical model for energy consumption, and is involved in the change of DTE when using a heating system or air conditioning system, considering not only driving energy but also Low DC converter(LDC). Unlike physical models, driving mileage-based DTE models are data-based, so they are made in consideration of driving patterns and various driving situations. This study is conducted with the same strategy as the latter. The data-based approach requires high-quality data and feature engineering for each variable. However, there are many features that are not related to the data, so preprocessing is needed.

III. DEFINITION OF DRIVABLE MILEAGE AND DISTANCE TO EMPTY

There are various factors that cause customers who have internal combustion engines or hybrid vehicles to purchase electric vehicles. According to Castro's report [5], the price of the vehicle, the charging time, and All Electric Range (AER). Batteries and motors account for 25 percent of the price of electric vehicles. Several battery companies are trying to improve battery prices and efficiency. Customers are very interested in charging. Compared to the fast refueling time of the internal combustion engine, the charging time of 30 minutes is also very long. AER is also an important purchasing factor. Range anxiety is the term that electric vehicle owners are afraid of discharge. Each electric vehicle developer is launching vehicles with a wider driving range after charging, and Mercedes-Benz introduced an electric vehicle that can travel 1,000 kilometers after charging once at CES 2021(Consumer Electronics Show 2021). Although it reduced range anxiety about charging by improving AER, there is a strong interest in whether it is possible to drive under the current SOC of battery. Electric vehicle customers who need to charge at public charging stations feel range anxiety more unlike the customers who charge at home or companies. Charging plans will be made with driving conditions and remaining mileage. The remaining mileage is defined as Distance To Empty (DTE). The remaining mileage for each electric vehicle is estimated based on distance or energy. Hyundai/Kia electric vehicles estimate the remaining mileage based on energy. It is made based on the remaining energy (kWh) of the battery and the estimated fuel economy (km/kWh) of the vehicle. The remaining energy of the battery does not decrease rapidly and can be predicted relatively, but the fuel efficiency of the vehicle has rapidly changing factors such as driving pattern and driving situation, and it is difficult to predict uncertain situations in the near future. Fig. 2 shows DTE equation, the history of the past fuel economy (km/kWh) is stored as a learning value in EPROM(Electrically Erasable Programmable Read-Only Memory) and the estimated fuel economy is calculated by blending the fuel economy in the current operation in Fig. 3. The remaining mileage is estimated with the determined fuel economy. This is a driving DTE.

Additionally, energy consumption by ancillary devices like air conditioning, electrical load, leakage current, etc. is also used. DTE according to air conditioning system is called air conditioning DTE. The entire remaining driving distance consists of a driving DTE and an air conditioning DTE. This study deals with driving DTE only.

IV. OUR WORK

A. DTE Model Reflecting Driving Patterns and Bayesian Probability

The existing basic driving DTE was determined by calculating the fuel economy with the current and past fuel economy, and calculating the fuel economy with the residual energy. It is implemented based on a physical model for energy consumption. Since the driving pattern also affects the driving DTE, if the driver changes, the DTE value from model changes, resulting in an error. Changes in drivers, friction changes in accessories, and battery performance affect the driving DTE. So, it is not easy to implement and integrate physical model. This study implemented driving DTE as a Bayesian probability linear regression model based on a linear regression model to improve the performance.

B. Drive DTE improvement recognizing driver’s past driving patterns

Selecting a destination in the navigation will inform you of the recommended route, current traffic conditions, and expected arrival time. The current driving DTE receives driving and route information from the navigation and...
corrects the fuel economy. Based on the information on the road type for each section between the departure and the destination, the average speed in the road type is reflected in the fuel economy calculation. This improved DTE is that the driver must select a destination in the navigation and have path data to the selected destination for correction. This study corrects the driving DTE with road type and driving pattern with past driving information and current departure information without navigation information selected by the current customer.

C. Overview of Machine Learning Model Development Process

This study is implemented for two models in driving DTE. These are the driving DTE model reflecting the Bayesian linear model and the driving pattern cluster model personalized. Unlike the existing DTE model based on the physical model, these models are implemented as machine learning models based on data. Fig. 4 shows the development process of a data based machine learning model.

For driving DTE machine learning, dataset of a EV6 electric vehicle for development was chosen. Data was collected for 4 months, and air conditioning system was not used only for driving DTE. Data were collected 60 times through INCA and ETAS data acquisition and mileage was 400 km. In the case of the driving pattern clustering model, driving data of several drivers were used because it contained the characteristics of each driver. The target vehicles were EV6 vehicles, and data were collected for 6 months. The algorithm was implemented using python and machine learning packages for the steps excluding the electric vehicle driving data acquisition stage and the vehicle verification stage. For the acquired data, the characteristics of the data were checked through EDA(Exploratory Data Analysis), and missing value processing, variable processing, and derived variable generation were performed. Supervised learning or unsupervised learning was defined according to the necessary purpose. The driving DTE model was supervised learning and consisted of a regression model, and the personalization model of driving pattern was from unsupervised learning. For performance comparison between machine learning models, Mean Square Error(MSE), Mean Absolute Error(MAE) and R² were used. In the case of driving DTE, the performance was confirmed through vehicle verification, and the personalization model was verified with separately acquired drivers’ driving data.

D. Development of a regression-based drive DTE model

The existing driving DTE model is a physical model based on energy consumption and is composed of the energy consumption and the fuel economy (km/kWh) from the battery. This model does not reflect real-time driving patterns or traffic conditions. This study was implemented by reflecting traffic conditions and driving patterns in a regression model for the remaining mileage. Input variables include vehicle speed, acceleration, mileage, accelerator pedal, brake, lateral/longitudinal acceleration, SOC (State of Charge) and SOC rate. The target variable is the driving mileage shown in DTE. Here, the driving mileage was changed to a form in which the value of the accumulated distance actually traveled was converted inversely and decreased over time. At the end of the driving, the DTE value becomes 0. Since the target DTE made up of only short-distance driving has a short mileage, short-distance driving may cause overfitting. Therefore, in the case of Driving Cycles (DC) that were not charged between each DC, data was attached to make one DC. It is assumed that the energy consumption ratio based on the same operation is not different depending on the start amount of SOC. Each DC was stacked with one data frame to generate data for learning and verification.

Random Forest, LightGBM (Light Gradient Boosting Machine), decision tree, and linear regression were candidates, Random Forest and LightGBM are ensemble models. The decision tree basically has an if-then structure and implements a learning model by extending the branches of the tree. Ensemble models are structured by attaching several sub-models as if trees were gathered to create a forest. When considering the model, a model that can be embedded in the controller is chosen except for Extreme Gradient Boosting (XGBoost), CATBoost, and Support Vector Regressor (SVR), which have excellent regression performance but high complexity of the model and long computing time.

For the data, train data and validation data were divided into a ratio of 7:3, and learning and verification were performed. GridsearchCV was used for hyper parameter tuning during learning, and the train data were set to five k-fold cross validations.

The performance of the models was evaluated by R² (1), MAE (2), MSE (3). A model with excellent performance was selected.

\[R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}\] (1)

\[\text{MAE} = \frac{1}{N} \sum_{i=1}^{n} |y_i - \hat{y}_i| \] (2)

\[\text{MSE} = \frac{1}{N} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \] (3)

Where \(y_i\) is target value, \(\hat{y}_i\) is predicted value, and \(\bar{y}\) is average value.

The regression model is the DTE model driven at development vehicle. However, if the customer continues to drive, the initial predictive performance may be degraded due to environmental factors, degradation of battery performance,
or changes in friction of the vehicle. If the degradation of the performance falls below the threshold by periodically monitoring the performance of the model, the model must be learned again. If the model learns with the entire dataset when re-learning in the vehicle's controller, it may cause the problem of insufficient resources for storing data. On the contrary, if the model learns with only recent data, the small data cause overfitting. To compensate for this, Bayesian linear regression model is proposed. Bayes theory is a theorem that represents the relationship between the prior probability and the posterior probability of two probability variables [6]. The existing linear regression model is defined as a prior probability distribution, and a new linear regression model is defined as a posterior probability distribution using the probability distribution of a group of input data. The regression model of the prior probability is the maximum likelihood estimation (MLE) that finds the optimal parameter (4).

\[
p(H|D) = \frac{p(D|H)p(H)}{p(D)} \tag{4}
\]

Where \(p(H)\) is prior probability model, \(p(D)\) is probability distribution of new data, \(p(D|H)\) is likelihood, and \(p(H|D)\) is the posterior probability model when new data are given. \(D\) is data, and \(H\) is hypothesis.

When new data \(D\) comes in, new parameters is determined with a high probability of \(p(D|H)\) and the existing parameter with this gets updated through Fig. 5.

As time passes, new data is collected, and the error increases when the performance of the existing model is checked with data created by new influences. So, to improve performance, the Bayesian model updates the parameters with new data.

### E. The Personalization model of the driving pattern cluster

Most of drivers have constant destinations. For example, commuting for work or school commuting for children follows regular routes. The mileage is similar and the destination is the same. This constant driving pattern does not always turn on the navigation. If the past driving history is stored and the driving starts at a certain date/time, it is recognized as the same driving as before, and the DTE can be calculated with the driving pattern. The key for this algorithm is to look at the driving history and create clusters through machine learning. Cluster is a method of unsupervised learning to divide clusters. K-means is a typical clustering model. Cluster models such as K-means must define the number of clusters in advance. However, the number of clusters showing constant driving varies individually. Therefore, this algorithm used a density-based DBSCAN method [8]. It automatically determines the number of clusters by looking at the density. Clusters were generated by vectorising driving time, date, day of the week, weekend/week, and mileage. An algorithm for personalizing a driving pattern by generating a cluster is as follows Fig. 6.

Driving data was used for dataset for learning and verification. Due to the characteristics of consequent driving data, it is acquired as statistical compressed data per drive cycle (DC). This is why it is impossible to analyze thousands of driving data as time series data directly. The start time of driving, the mileage at one DC, and additional variables were obtained by receiving them as statistical values.

The personalization model of driving pattern converts the time vector into x and y coordinates. If time is used as a vector as it is, 23:59 and 00:01 become very long distant vectors. However, if it is vectorized into x and y coordinates using sine and cosine like an analog clock, 23:59 and 00:01 are very close. Clusters were created using time vectorization at the time of start, day of the week, weekly/weekend, and mileage at one run as vectors. As for the cluster, DBSCAN, a density-based cluster model, was selected.

### V. RESULT

#### A. Result of Regression Models of Driving DTE

In the case of the driving DTE model, a linear regression model was selected as the basic model. It was selected because it had linearity due to the characteristics of the data and is simple compared to the ensemble model of the tree structure, and has good interpretation and excellent performance. Evaluation criteria such as \(R^2\), MAE, and MSE
TABLE I. PERFORMANCE OF EACH MODEL

<table>
<thead>
<tr>
<th>Performance</th>
<th>R² score</th>
<th>MAE</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear Regressor</td>
<td>0.97</td>
<td>4.77</td>
<td>35</td>
</tr>
<tr>
<td>LightGBM</td>
<td>0.66</td>
<td>7.8</td>
<td>150</td>
</tr>
<tr>
<td>Random Forest</td>
<td>0.92</td>
<td>4.5</td>
<td>53</td>
</tr>
<tr>
<td>reference</td>
<td>0.84</td>
<td>5.32</td>
<td>47</td>
</tr>
</tbody>
</table>

was selected. The linear regression was compared with tree-structured ensemble models like LightGBM and Random Forest. The comparison results are shown in Table 1, and Fig.7. Where “pred” is the estimated profile with linear regression, “true” is the modified actual driving mileage profile, and “origin” is the existing DTE profile.

It can be seen that the linear regression model performs better than the existing DTE model. In addition, the ML model has to be embedded in ECU, but complex models occupy resources and calculate late in ECU. Because of this, linear model has benefit rather than other tree based models.

B. Result of Bayesian regression model for Driving DTE

In order to implement the Bayesian regression model, new data had to be obtained. Additional driving data were collected, and the total mileage was 200 km. The performance was evaluated with new data for the existing regression model, and the Bayesian regression model was evaluated in the same way. The evaluation results are shown in Table 2.

"Before 200km with linear regression" is the performance evaluation for the regression model initially, and "After 200km with linear regression" is the result of the performance change for the same model after the 200 km drive. The last "After 200km with Bayesian regression" is an evaluation of the results predicted in the Bayesian regression model.

C. Results of personalization of driving pattern clusters

Personalization model of the driving pattern cluster stores the past driving history and uses the driving history if there is the similar driving history during the current driving. The driving pattern cluster was generated by DBSCAN. DBSCAN is density based clustering model, but the number of epsilon and samples must be determined [7]. The mileage of 500 m less and the driving time of 30 minutes less were removed. The distance of vectors in a cluster was limited to limit the number of epsilons and samples. Each individual may have a different number of epsilon and samples because each individual may only have a scheduled drive to commute work or school. The cluster results are shown in Fig. 8 and Fig. 9. Fig. 8 is a table of clustered driving, and Fig. 9 is visualized by the driving start time of driving and the distance of driving. The dow is day of week. This customer drives on weekdays between 7:30 and 7:40, and travels about 10km. And in the figure on the right, X and Y coordinates are vector values of time, and distance is the value of normalizing the distance. It can be seen that clustered data are gathered together.

The collected data for one driver were from one city. The cluster is for only one driver, and many cluster models is for many drivers individually.

Fig. 10 shows the cumulated counts by day of the week for a month as drive pattern of certain driver, and it can be seen that the clusters of 1 to 4 are evenly distributed by day of the week, and the cluster of 5 is gathered only on Saturdays in the daily clusters for a month. The cluster model is DBSCAN. Weekend in Korea is two days.

It can also be seen that in the case of Sundays, they do not go to a specially designated place.

One of clusters says that the customer will drive in this time with the start value of the drive. For example, if you drove at 7:38, you would be included in the cluster of driving rules as shown in the fig. 8. The cluster is determined by the Euclidean distance method as to which cluster the current

TABLE II. PERFORMANCE BETWEEN LINEAR REGRESSION AND BAYESIAN REGRESSION

<table>
<thead>
<tr>
<th>Performance</th>
<th>Before 200km with linear regression</th>
<th>After 200km with linear regression</th>
<th>After 200km with Bayesian regression</th>
</tr>
</thead>
<tbody>
<tr>
<td>R²</td>
<td>0.97</td>
<td>0.91</td>
<td>0.95</td>
</tr>
<tr>
<td>MAE</td>
<td>4.77</td>
<td>10.1</td>
<td>5.1</td>
</tr>
<tr>
<td>MSE</td>
<td>35</td>
<td>47</td>
<td>39</td>
</tr>
</tbody>
</table>
Although deep learning model has better performance [8], it is difficult to embed in limited controller hardware. Model compression is necessary to mount deep learning with tons of parameters. Therefore, python code has to be converted into C++ code. In addition, although, Deep Learning and tree based machine learning have good performance, it is risky to imbed AI model into ECU because it is difficult to revise model in ECU after mass production. In vehicle mass production, interpretability is very important issue. The next study is to improve interpretability for revising and to convert python code into C++ for fast computing.

Through this improvement in performance, predicting the remaining driving mileage of the electric vehicle provides convenience as the driver can plan in advance when to charge the destination to which they want to go. It is expected that the customer will be able to clearly establish a charging plan and increase product satisfaction by checking more accurate DTE with their driving pattern and driving history.

VI. CONCLUSION

Since electric vehicles have the disadvantage that they take longer to charge than internal combustion engine vehicles, drivers should know the current drivable mileage and plan the driving path or mileage in advance. This study achieved a more accurate prediction of the Distance To Empty(DTE). Although deep learning model has better performance [8], it is difficult to embed in limited controller hardware. Model compression is necessary to mount deep learning with tons of parameters. However, this linear regression model has less than 10 parameters, and performance improvement is not parameter tuning, but derived variables are generated, so it is possible to define in advance, and explain which variables are of high importance. It was confirmed that the verification performance was superior compared to other ensemble models. Linear regression models with good portability, model handling, and explanatory power were selected for the controller.

Personalization model of driving pattern cluster is possible to predict the remaining mileage using the past driving history, without navigation information, and it can also be used to calculate the fuel economy for each road type in the remaining mileage based on current energy. The performance of the customer's driving pattern cluster personalization model was confirmed through supervised learning.

Through this study, the future task is to implant the model into a controller equipped with a high-performance process and verify it in the vehicle. A controller equipped with a high-performance process is what can be operated by implanting a machine learning or deep learning model, and this machine learning model is expected to be portable. However, python code takes a lot of time to operate. Therefore, python code has to be converted into C++ code. In addition, although, Deep Learning and tree based machine learning have good performance, it is risky to imbed AI model into ECU because it is difficult to revise model in ECU after mass production. In vehicle mass production, interpretability is very important issue. The next study is to improve interpretability for revising and to convert python code into C++ for fast computing.

Through this improvement in performance, predicting the remaining driving mileage of the electric vehicle provides convenience as the driver can plan in advance when to charge the destination to which they want to go. It is expected that the customer will be able to clearly establish a charging plan and increase product satisfaction by checking more accurate DTE with their driving pattern and driving history.

REFERENCES